ASJ International Journal of Advances in Medical Sciences and Biotechnology (IJAMSB)
Vol. 3(1) 10 June, 2017, Pp. 50-59
www.academiascholarlyjournal.org/ijamsb/index_ijamsb.htm
Indexed In: Directory of Research Journals Indexing - http://www.drji.org
Also Available@; www.archive.org/details/Onoh_and_Nwaogazie
Open access

Full Length Research

Obesity Classification Based On Agglomerative Hierarchical Clustering

Charles C. Onoh[1] and Ify L. Nwaogazie[1*]

[1]Centre for Occupational Health, Safety and Environment, University of Port Harcourt, Nigeria.

Accepted June, 05, 2017

Obesity is the excessive accumulation of fat in the body which adversely affects the health and well-being of the individual. It is a chronic and non-communicable disorder that poses socio-cultural, psychological, clinical and public health challenges. The aim of this study is to apply Agglomerative Hierarchical Clustering (AHC) to classify obesity and to develop a model employing Logistic regression analysis for the prediction of obesity taking advantage of the relationship between Body Mass Index (BMI), Age, Waist Circumference (WC), High-Density Lipoprotein (HDL)-cholesterol and Low-Density Lipoprotein (LDL)-cholesterol. This Study was a work–site based cross sectional study carried out on one hundred and twenty (120) workers at Judiciary Service Commission, Owerri, Imo State, Nigeria. The Questionnaire was designed to address the background information of the respondents with respect to gender, age, job title, department and address. The respondents were anthropometrically examined and their lipid profile was estimated using the enzymatic colorimetric method. Data were analysed using the Shapiro-wilks test of normality, Agglomerative Hierarchy Cluster (AHC) analysis and Logistic regression analysis. These analyses were facilitated using XLSTAT 2016 statistical tool. On the application of the Agglomerative Hierarchical Cluster Analysis obesity was classified into Clusters 1, 2 and 3 with the majority of the obesed respondents being in Cluster 1. The respondents in Cluster 1 belonged to the obesity class of overweight, while respondents in Cluster 2 are of normal weight and finally respondents in Cluster 3 belonged to obese class 1. A predictive model was developed based on Logistic regression analysis which showed a strong positive correlation between obesity and HDL-cholesterol. The high profile of cardiovascular risks identified in the study could be addressed through the provision of occupational health services of which the ultimate goal should be the maintenance of urgent comprehensive health surveillance.

Keywords: Obesity classification, BMI, Age, Waist circumference and AHC.

INTRODUCTION

Obesity is the excessive accumulation of fat in the body as to adversely affect the health and well-being of the individual (Siminnialayi et al., 2008). It is a chronic and non-communicable disorder that poses socio-cultural, psychological, clinical and public health challenges. It is a risk factor for hypertension, diabetes, degenerative arthritis, myocardial infraction sleep apnoea, and various

***Corresponding Author's Email: ifynwaogazie@yahoo.com**

forms of cancer.

The global burden of obesity is more than 1.1billion. Worldwide obesity has more than doubled since 1980. In 2014, more than 1.9 billion adults were overweight out of which over 600 million were obesed. Obesity results from a complex interplay of environmental and genetic factors. It is associated with significant morbidity and mortality. It is also associated with emotional consequences, social stigmatization, increased risk of various medical diseases (Aronne, 2012). The aetiology of obesity is a constellation of factors, multifactorial involving the environment, nutrigenetic and behavioural factors. Risk factors include sedentary lifestyle and interplay of some hormones like leptin and irisin.

Body Mass Index (BMI) is a method employed in measuring obesity. Sometimes it bears the name Quetelet Index (QI) because it was developed by the Belgian statistician and anthropometrist Adolphe Quetelet (Quetelet, 1871). BMI is calculated by dividing the weight in kilograms by the square of the height in metres (m^2). It is an index of defining and classifying obesity and has replaced percentage ideal body weight as a criterion for assessing obesity. BMI in kg/m^2 of ≤ 18.5; 18.5 – 24.9; 25.0 – 29.9; 30.0-34.9; 35.0-39.9; ≥ 40.0 are underweight, normal weight, overweight, obese class I, obese class II and obese class III, respectively (Grima and Dixon, 2013).

The aim of this study is to apply Analytical Hierarchical Clustering (AHC) to classify obesity and to develop a model applying logistic regression analysis for the prediction of obesity taking advantage of the relationship between BMI, age, Waist Circumference (WC), High-Density Lipoprotein (HDL)-cholesterol and Low-Density Lipoprotein (LDL)-cholesterol.

METHODOLOGY

Study Area

This Study was carried out in Owerri, the capital and largest city in Imo state. Imo state is one of the States in South Eastern region of Nigeria. It is bounded by the Otamiri and Nworie rivers on the east and south, respectively (Onoh and Nwaogazie, 2016). It lies within longitude 7.63 and latitude 5.48. The people are mainly civil servants and traders. The town is filled with numerous restaurants and eating places with bars. The town bubbles with night life all days of the week.

Population of the Study

The study population was Judiciary Service Commission, Orlu Road, Owerri. The population comprised of civil servants in Judiciary Service Commission, Owerri. The Commission's registers were used for the population/sampling frame. One hundred and twenty (120) participants were involved between April and September, 2016.

Inclusion Criteria

The criteria included bonafide workers in Judiciary Service Commission, Owerri who are 18 years and above. They must have signed the informed written consent to participate in the study.

Exclusion Criteria

Excluded from this study were civil servants with physical deformity. Equally excluded were the pregnant women and elderly workers.

Sample Size Estimation

Sample size was estimated using the prevalence formula (see Equation 1).

$$N = \frac{Z^2 P(1-P)}{T^2} \quad (1)$$

Where N represents the sample size, T = tolerance error (0.05); P = prevalence of previous study (Iloh et al., 2010, 2011 and 2013) and Z = 1.96, which is the level of significance and corresponds to 95% confidence level. Evaluating Equation (1) yields:

$$N = \frac{1.96^2 \, X \, 0.075(1-0.075)}{0.05^2} = 106.6 \approx 107$$

The sample size was calculated as 107, adopting 10% as attrition rate yielded value of 120. Thus, 120 workers were recruited for the study.

Methods of Data Collection

The methods adopted for data collection were questionnaires and focused group discussions. It is common knowledge that in most public health research in literature, questionnaires and focused group discussions form the basis of data collection.

Most of the data appear as categorical variables involving multiple categorical variables.

Training of Research Assistants

The simple random sampling technique was applied with regard to data collection wherein five nurses were trained on anthropometric measurements in this study.

Questionnaires

The questionnaires were designed to capture the socio-demographic variables such as age, sex, occupation, weight and height.

Method of Data Analysis

The method of data analysis applied on the collected data with respect to this study includes the Shapiro-wilks test of normality, AHC analysis and logistic regression analysis which were carried out using XLSTAT 2016 statistical software tool. The test of normality was to determine the data type whether parametric or non-parametric and thus, aid the choice of regression option. Shapiro-wilks *W* test of normality has been found to be the most powerful test in most of the situations (Razali and Wah, 2011). It employs the ratio of two estimates of the variance of a normal distribution based on a random sample of *n* observations. *W* is roughly a measure of the straightness of the normal quantile-quantile plot. Hence, the closer *W* is to one, the more normal the sample is.

AHC was applied on the collected data set in order to classify them to obtain the various classification of obesity with respect to the study population. AHC is an important and inherent process in unsupervised machine learning. It begins with each variable signifying an individual cluster. These are then subsequently merged according to their similarity. Initially, the two most similar clusters (usually those with the smallest distance between them) are merged to form a new cluster at the bottom of the hierarchy. In the next step, another pair of clusters is merged and linked to a higher level of the hierarchy, and so on (Jones et al., 2001). The similarity proximity type employed in this study was with respect to Kendal correlation coefficient and the agglomerative method used was the un-weighted pair-group average approach. Cornell et al. (2007) applied AHC to identify clinically relevant multimorbidity groups. Also, Deliu et al. (2016) applied clustering approach into identification of asthma subtypes. Furthermore, Arnob et al. (2017) applied AHC for the classification of Phylogeny in order to identify the centroid representative genes.

Furthermore, logistic regression analysis was carried out on the collected data set to model at a confidence interval of 95%. Logistic Regression is a useful mathematical modeling approach for analyzing the relationship between data that includes categorical response variable such as the presence or absence of a disease to a dichotomous dependent variable (Bererud, 1996; Anderson et al., 2003; Kleinbaum and Klein, 2010). Equation (2) presents the logistic function $f(z)$ which describes the mathematical form on which the logistic model is based. This function ranges between 0 and 1, thus, the model is designed to describe a probability which is always some number between 0 and 1. In epidemiologic terms, such a probability gives the risk of an individual getting a disease. For multiple logistic regression with independent variable $x_1, x_2, x_3, \ldots x_n$ (as the case of this study), Equation (2) can be further presented by Equation (3). Equation (3) can further be expressed in terms of its logit value as Equation (4) where p with respect to this study is the probability of a respondent being obese. For the regression modeling, BMI value was taken as the dependent variable while the age, waist circumference, HDL, and LDL were the independent variables represented by x_1, x_2, $x_3$and x_4, respectively. The collected data set with respect to BMI, HDL, WC, and LDL with the exception of age data were converted to binary data set with BMI > 25 kg/m^2, WC > 82 cm, HDL < 50, and LDL >100 equal to 1(probability of obesity) while BMI < 25 kg/m^2, WC < 82 cm, HDL > 50, and LDL <100 equal to zero (probability of not being obesity) (Alberti et al., 2006).

$$f(z) = \frac{1}{1+e^{-z}} \tag{2}$$

$$P(y \mid x_1, x_2, x_3, x_4) = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3}} \tag{3}$$

$$\text{logit (p)} = \ln\left\{\frac{\Pr(y=1)}{\Pr(y=0)}\right\} = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 \tag{4}$$

Table1. Summary of Shapiro-wilks Test of Normality.

Variable\Test	Shapiro-Wilk	Data Type
BMI[kg/m^2]	< 0.0001	**Non-parametric**
AGE [yrs]	< 0.0001	**Non-parametric**
WC [cm]	0.0311	**Non-parametric**
HDL	< 0.0001	**Non-parametric**
LDL	< 0.0001	**Non-parametric**

RESULTS

On applying Shapiro-wilks test of normality on the collected data set, Table 1 presents a summary of the normality test (see Appendix A). While Figures 1a – 1e present the normal p-p plot of the resultant theoretical distribution against the empirical distribution of the various variables in the data sets. Table 2 is a summary of the collected data set with respect to the variables.

On application of AHC, Figure 2 presents the resultant dendrogram. Table 3 presents the summary of the analysis with respect to the various resultant classes (clusters). Table 4 presents the central object which defines the various classes with respect to this study. The attained output from further analysis of the collected data set using logistic analysis is presented in Table 5. Also, Table 6 presents the standardize coefficient of the variables with respect to the regression analysis while Equation (5) present the resultant logistic regression model.

$$\text{logit (BMI)} = \ln\left\{\frac{\Pr(y=1)}{\Pr(y=0)}\right\} = -0.55 + 2.33WC + 1.15LDL \quad (5)$$

DISCUSSION

The collected data with respect to this study were subjected to normality test to determine the type of data analysis. The essence of determining the nature of data is to aid the choice of statistical analysis to be applied in this study. From Table 1, on applying Shapiro-wilks test of normality, all the variables with respect to the collected data set were not normally distributed (see Appendix A). This can also be seen in Figure 1 which presents the normal probability plot output of the collected data set. The collected data set (blue-dots) do not follow a linear pattern (black dotted line) but seem to scatter around it. Hence the nature of the data was assumed non-parametric (Razali and Wah, 2011) with respect to the present study.

The output from subjecting the collected data set to AHC presented three major clusters or classes (C1, C2, and C3) as shown in Figure 2. It is interesting to note that classes 2 and 3 are closely related (see Figure 2). Class 1 comprised of fifty (50) respondents, while classes 2 and 3 were of forty-two (42) and twenty-eight (28) respondents, respectively. From Table 4, the female gender was the central object around which the various classes or cluster were defined having their BMI values as 25.39, 24.74 and 34.62, respectively. From the World Health Organization standard for Obesity classification, respondents in class 1 belong to the class of obesity, that is, over-weight, while class 2 respondents are of normal weight and finally, respondents in class 3 belong to the obese class 1 group (Grima and Dixon, 2013). The preponderance of the female gender implicated in the obesity class may be due to hormonal interplay, multiparity, sociocultural and traditional practices. The female gender is more sedentary in nature and more importantly nutrigenetic factors are implicated (Amoah, 2003). Furthermore, the similarity seen between classes 2 and 3 as presented by the dendrogram could be attributed to the closeness of their waist circumference (WC) and High-Density Lipoprotein (HDL)-cholesterol centroidal values (see Table 4).

The likelihood of the resultant model with respect to the probability of obesity was equal to 93.14 (see Table 6). Also, from the resultant logistic model (see Equation 5 and Table 6), WC statistically has no significant effect on the probability of obesity (p value < 0.0001) unlike LDL which has a significant effect on the probability of obesity (p value = 0.0732). Age, HDL both had coefficients equal to zero, at WC equal to zero (being not significant). The odds in favor of obesity for a unit LDL increases by a multiplicative factor of $3.16\left(e^{1.15}\right)$. The findings of this study are in agreement with the works of Agu et al. (2015), Omotoye and Fadupin, (2016) with LDL having a positive and significant correlation with obesity (that is BMI).

CONCLUSION

Based on the result of this study, the following

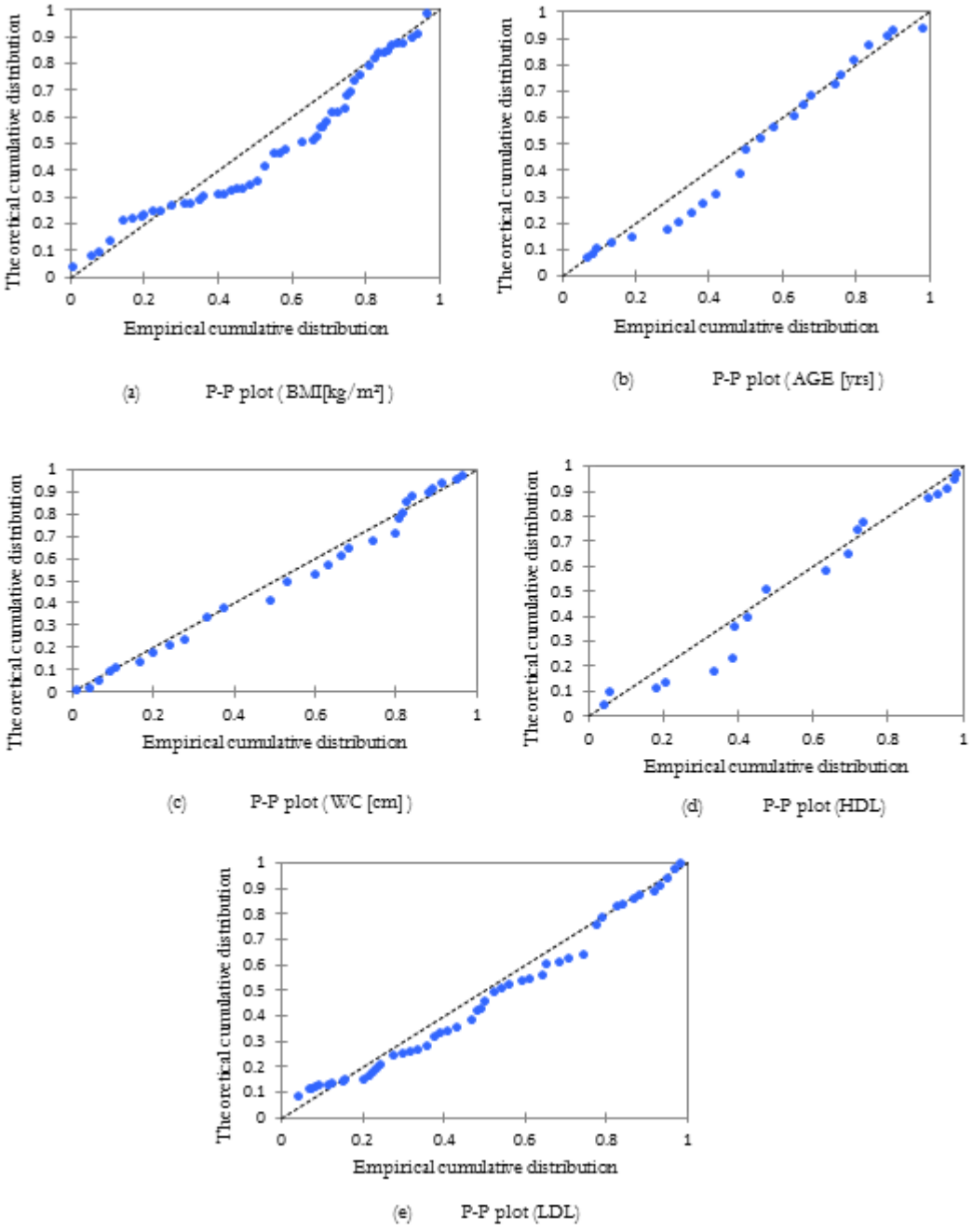

Figure 1. Resultant p-p Plot from normality test of the variables in the collected data set.

i. This study has established that Obesity could be classified into Clusters 1, 2 and 3. The respondents in Cluster 1 belonged to the obesity class of overweight, while respondents in Cluster 2 are of normal weight and in Cluster 3 belonged to obese class I.

ii. The resultant Logistic model showed that LDL-cholesterol has a significant effect on the probability of obesity. It has a

Table 2. Summary of the collected data set.

Variable	Minimum	Maximum	Mean	Std. deviation
BMI[kg/m2]	19.5804	44.7368	28.5372	5.2208
AGE [yrs]	24.0000	58.0000	37.5250	9.1990
WC [cm]	71.0000	114.0000	93.1917	10.1603
HDL	40.0000	82.0000	57.7917	10.6990
LDL	36.0000	228.0000	92.2417	40.5118

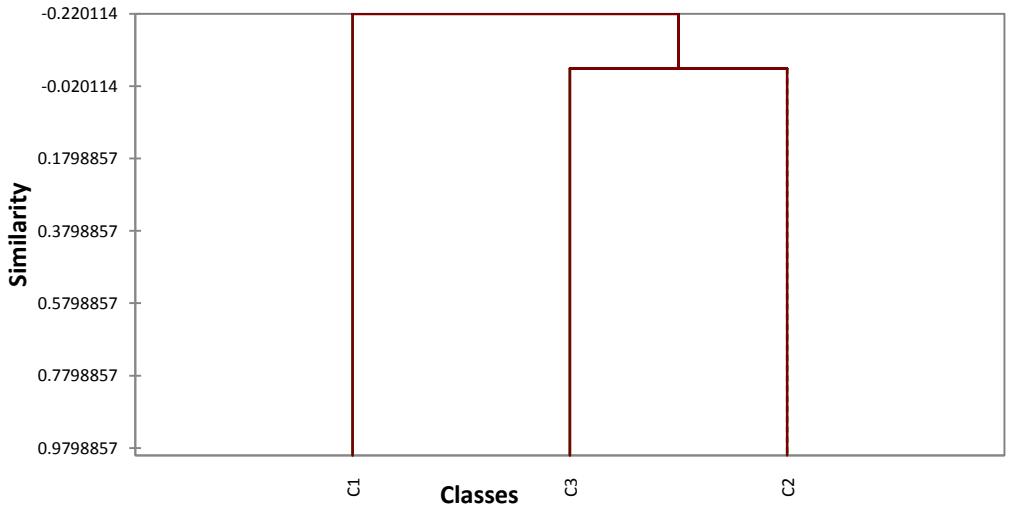

Figure 2. Resultant dendrogram.

Table 3. Summary of Resultant AHC of collected data set.

Class	1	2	3
Objects	50	42	28
Sum of weights	50	42	28
Within-class variance	1872.7694	690.2601	1226.9789
Minimum distance to centroid	8.8405	9.6185	10.8999
Average distance to centroid	35.6587	24.3932	31.6718
Maximum distance to centroid	111.3753	45.6384	50.4252

significantly positive correlation with obesity.

iii. A Model was developed based on Logistic Regression Analysis which showed a strong positive correlation between obesity and HDL-cholesterol.

iv. Obesity was a predominant cardiovascular risk identified and the female gender showed a preponderance of obesity morbidity.

Table 4. Central objects with respect to the resultant classes.

Class	BMI [kg/m^2]	AGE [yrs]	WC [cm]	HDL	LDL
1 (F)	25.3906	32.0000	86.0000	45.0000	120.0000
2 (F)	24.7449	35.0000	101.0000	66.0000	60.0000
3 (F)	34.6154	24.0000	98.0000	58.0000	94.6000

Table 5. Goodness of fit statistics (Variable BMI [kg/m^2]) from Logistic Regression.

Statistic	Independent	Full
Observations	120	120
Sum of weights	120.0000	120.0000
DF	119	117
-2 Log(Likelihood)	114.3385	93.1487
R²(McFadden)	0.0000	0.1853
R²(Cox and Snell)	0.0000	0.1619
R²(Nagelkerke)	0.0000	0.2635
AIC	116.3385	99.1487
SBC	119.1260	107.5112
Iterations	0	6

Table 6. Standardized coefficients (Variable BMI[kg/m^2]).

Source	Value	Standard error	Wald Chi-Square	Pr> Chi²	Wald Lower bound (95%)	Wald Upper bound (95%)	Odds ratio	Odds ratio Lower bound (95%)	Odds ratio Upper bound (95%)
Intercept	-0.5500	1.2929	0.1809	0.6706	-3.0839	1.9840			
AGE [yrs]	0.0000	0.0000							
WC [cm]	2.3260	0.5709	16.6006	< 0.0001	1.2071	3.4449	10.2365	3.3437	31.3388
HDL	0.0000	0.0000							
LDL	1.1527	0.6434	3.2095	0.0732	-0.1084	2.4139	3.1669	0.8973	11.1773

RECOMMENDATION

Based on the findings from this study the following recommendations were made:

i. The high profile of cardiovascular risks identified in the study could be addressed through the provision of occupational health services.

ii. There is need for preventive programmes like moderate endurance aerobic exercise and medical nutrition therapy to address the issue of obesity.

iii. Practitioners in the medical and managerial fields could apply Agglomerative Hierarchical Cluster Analysis in making decisions regarding diagnoses, interventions or educational campaigns among populations at risk. HDL-cholesterol estimation could also be employed in the assessment of obesity.

REFERENCES

Agu CE, Emeribe AU, Idris AN, Effa AF, Adamu B and Alphonsus EU (2015). Evaluation of fasting lipid profile and glycated hemoglobin in obese subjects at University of Calabar teaching hospital,

Nigeria. Int. J. Biomed. Res., 6(3): 200-209.
Alberti KGMM, Zimmet P and Shaw J (2006).Metabolic syndrome—a new world-wide definition.A Consensus Statement from the International Diabetes Federation. Diabetic Med., 23: 469–480.
Amoah AG (2003). Socio-demographic variation in obesity among Ghanaian adults. Public Health Nutr; 6:751-757.
Anderson RP, Ruyun J and Grunkemeier GL (2003). The statistician Page-Loistic Regression Analysis in Clinical Reports: An Introduction. Ann. Thorac. Surg., 75:753-757.
Arnob RI, Sony KR, Mottalib MA, Akter L and Kushol R (2017). Advances Agglomerative Clustering Technique for Phylogenetic Classification. 1st International Conference on Engineering Research and Practice, Dhaka, Bangladesh. Pp. 13-18
Aronne LJ (2012). Classification of obesity and assessment of obesity- Related health risks. Obesity J., 10(S12): 105S-115S.
Bererud WA (1996): Introduction to Logistic Regression Models with Worked Forestry Examples – Biometric Information Handbook No.7. Res. Br., B.C. Min. For.,Victoria, B.C. Work. Pap. 26/1996.
Cornell JE, Pugh JA, Williams JW, Kazis L, Zeber J, Pederson T, Montgomery KA., Nokl PH (2007): Multibobidity Clusters: Clustering Binary Data from Multimobidity Clusters: Clustering Binary data from a Large Administrative Medical Database. Appl. Multivariate Res., 12(3) 163-182.
Deliu M, Sperrin M, Belgrave D and Custovic A (2016): Identification of Ashman Subtypes using Clustering Methodologies. PulmTher 2: 19 – 41. DOI 10.1007/s41030-016-0017-z.
Grima M and Dixon BN (2013).Recommendations for management of general practice and beyond obesity. 42(8): 532-541
Iloh GU, Amadi AN and Nwankwo BO (2010).Obesity in adult Nigerians.A study of its prevalence and common primary co-morbidities in semi-urban Mission General Hospital in South-Eastern Nigeria. Niger: Med. 19: 459-66.
Iloh GU, Amadi AN, Nwankwo BO and Ugwu VC (2011). Obesity in adult Nigerians: A study of its patterns and common primary co-morbidity in a rural Mission General Hospital in Imo State, South-Eastern Nigeria. Nig. J. Clin. Pract. 14(2): 212-218.
Iloh GU, Ikwudimma AO and Obiegbu NP (2013). Obesity and its cardiometabolic. Morbidities among adult Nigerians in a primary care clinic of a Tertiary Hospital in South-Eastern, Nig. J. Family Med. and Primary Care. 2(1): 20-26.
Jones E, Oliphant T and Peterson P (2001). SciPy: Open Source Scientific Tools for Python, http://www.scipy.org
Kleinbaum DG, and Klein M (2010). Logistic Regression, Statistics for Biology and Health -A Self-Learning Text. Third Edition.Springer New York Dordrecht Heidelberg London. Springer Science + Business Media, LLC. ISBN: 978-1-4419-1741-6 e-ISBN: 978-1-4419-1742-3. DOI 10.1007/978-1-4419-1742-1743
Omotoye FE and Fadupin GT (2016): Effect of Body Mass Index on Lipid Profile of Type 2 Diabetic Patients at an Urban Tertiary Hospital in Nigeria. IOSR Journal of Dental and Medical Sciences (IOSR-JDMS). 15(9): 65-70.
Onoh CC and Nwaogazie IL (2016). Workplace assessment of hypertension: Prevention and awareness in a food processing industry in Owerri, Nigeria. Int. J. Trop. Dis. and Health, 16(4): 1 – 11. Available at http://www.sciencedomain.org.
Quetelet LAJ (1871). Antopometricoumesure des différencesfacultés de l'Homme. Brussel: musqaurdt (*Anopometric of the Differences of Man. Brussel: musqaurdt*-English)
Razali NM and Wah YB (2011). Power Comparisions of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. J. Stat. Model. Anal., 2(1): 21-33
Siminnialayi IM, Emem-Chioma PC and Dapper DV (2008). The prevalence of obesity as indicated by BMI and waist circumference among Nigerian adult attending family medicine clinic as outpatients in Rivers state, Nig. J. Med., 17: 340-5

APPENDIX A

Table A1. Shapiro-Wilk test (BMI[kg/m2]).

W	0.9010
p-value (Two-tailed)	< 0.0001
Alpha	0.05

Test interpretation:

H_0: The variable from which the sample

was extracted follows a Normal distribution.
H_a: The variable from which the sample was extracted does not follow a Normal distribution.
As the computed p-value is lower than the significance level alpha=0.05, one should reject the null hypothesis H0, and accept the alternative hypothesis H_a.

The risk to reject the null hypothesis H0 while it is true is lower than 0.01%.

Table A2. Shapiro-Wilk test (AGE [yrs]).

W	0.9425
p-value (Two-tailed)	< 0.0001
Alpha	0.05

Test interpretation:

H_0: The variable from which the sample was extracted follows a Normal distribution.
H_a: The variable from which the sample was extracted does not follow a Normal distribution.
As the computed p-value is lower than the significance level alpha=0.05, one should reject the null hypothesis H0, and accept the alternative hypothesis H_a.

The risk to reject the null hypothesis H0 while it is true is lower than 0.01%

Table A3. Shapiro-Wilk test (WC [cm]).

W	0.9761
p-value (Two-tailed)	0.0311
Alpha	0.05

Test interpretation:

H_0: The variable from which the sample was extracted follows a Normal distribution.
H_a: The variable from which the sample was extracted does not follow a Normal distribution.
As the computed p-value is lower than the significance level alpha=0.05, one should reject the null hypothesis H0, and accept the alternative hypothesis H_a.

The risk to reject the null hypothesis H0 while it is true is lower than 3.11%.

Table A4. Shapiro-Wilk test (HDL).

W	0.9327
p-value (Two-tailed)	< 0.0001
Alpha	0.05

Test interpretation:

H_0: The variable from which the sample was extracted follows a Normal distribution.
H_a: The variable from which the sample was extracted does not follow a Normal distribution.
As the computed p-value is lower than the significance level alpha=0.05, one should reject the null hypothesis H0, and accept the alternative hypothesis H_a.

The risk to reject the null hypothesis H0 while it is true is lower than 0.01%.

Table A4. Shapiro-Wilk test (LDL).

W	0.9317
p-value (Two-tailed)	< 0.0001
Alpha	0.05

Test interpretation:

H_0: The variable from which the sample was extracted follows a Normal distribution.
H_a: The variable from which the

sample was extracted does not follow a Normal distribution.
As the computed p-value is lower than the significance level alpha=0.05, one should reject the null hypothesis H0, and accept the alternative hypothesis H_a.

The risk to reject the null hypothesis H0 while it is true is lower than 0.01%.